D. Cocchi and A. F. Di Narzo

A Bayesian Hierarchical Approach to Ensemble Weather Forecasting

Quaderni di Dipartimento

Serie Ricerche 2008, n. 5
ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche “Paolo Fortunati”

A Bayesian Hierarchical Approach to Ensemble Weather Forecasting

D. Cocchi and A. F. Di Narzo

Department of Statistics P. Fortunati,
University of Bologna, Italy

May 10, 2008

Abstract

In meteorology, the traditional approach to forecasting employs deterministic models mimicking atmospheric dynamics. Forecast uncertainty due to the partial knowledge of initial conditions is tackled by Ensemble Predictions Systems (EPS). Probabilistic forecasting is a relatively new approach which may properly account for all sources of uncertainty. In this work we propose a hierarchical Bayesian model which develops this idea and makes it possible to deal with an EPS with non-identifiable members using a suitable definition of the second level of the model. An application to Italian small-scale temperature data is shown.

Keywords. Ensemble Prediction System, hierarchical Bayesian model, predictive distribution, probabilistic forecast, verification rank histogram.

Acknowledgements The research leading to this paper has been funded by a 2006 grant (project n. 2006139812.003, sector: Economics and Statistics) for researches of national interest by the Italian Ministry of the University and Scientific and Technological Research.

We also would like to thank Emilia Romagna ARPA-SMR, for providing us the temperature dataset and in particular Chiara Marsigli, Tiziana Paccagnella and Stefano Tibaldi for the useful discussion.

1 Introduction

In meteorology, the typical approach to forecasting employs deterministic models mimicking atmospheric dynamics. Probabilistic forecasting is a relatively new approach attracting growing interest (see Gneiting et al., 2007). One technique currently used to tackle the problem of weather predictability is Ensemble Weather Forecasting. A prediction system based on this technique is called an Ensemble Prediction System (EPS). An EPS produces multiple weather forecasts by iterating forward random perturbations of a best estimate of initial conditions (Leith, 1974; Toth et al., 2001).

Weather forecasts obtained using an EPS can be synthesised to give one single value. Such a value may derive from an average, as in Raftery et al. 2005, or from a selection procedure, as in Roulston and Smith 2003, where the Best Member Dressing method is proposed. One important distinguishing feature of an EPS is the identifiability of its outputs. The competing models of Raftery et al. 2005 are indeed different forecasting models which may, however, agree upon a common evaluation. The model used in Roulston and Smith 2003 is completely different: in fact, this uses the EPS from the European Center for Medium-Range Weather Forecasts (ECMWF; Montani et al., 1996), where model outputs cannot be identified since they are replicas under perturbation of initial conditions. A number of different ways of treating EPS meteorological forecasts have been suggested.

The Bayesian Model Averaging (BMA; as in Hoeting et al., 1999) approach adopted by Raftery et al. 2005 focuses on one application to the University of Washington's multi-model EPS. It treats forecasts from deterministic models as inputs for a statistical model and can be extended to dynamical models. Of the forecasts f_k, each computed according to model k, there is a 'best' forecast, while the K models are

different and identifiable. Uncertainty about which model is the best one is naturally quantified by BMA. Denote as y the observed value, and each deterministic forecast as $\hat{f}_k$. Each deterministic forecast can be corrected for possible bias, for example by a linear transformation, thus giving the corrected forecast $f_k = a_k + b_k \hat{f}_k$. A conditional predictive distribution function $g_k(y|f_k, \theta)$ is associated to f_k and this function can be interpreted as the distribution of y conditional on f_k, given that f_k is the best forecast in the ensemble. The predictive distribution function as given by BMA is (cfr. Raftery et al., 2005, eq. (2)):

$$p(y|f_1, f_2, \ldots, f_K) = \sum_{k=1}^{K} w_k g_k(y|f_k, \theta) \tag{1}$$

where w_k is the probability that f_k is the best forecast, and is based on the predictive performance of the k-th dynamic model in a training set of deterministic forecasts and observed values. Weights w_k are model probabilities, and sum to 1. Estimation of parameter θ is performed by ML, with the EM algorithm. Expression (1) is referred to as Bayesian since it weighs likelihoods by evaluating the appropriateness of each model, which in this context can be seen as a 'state of the world'.

Roulston and Smith 2003 propose the Best Member Dressing method, and analyse the ECMWF EPS. Given an ensemble forecast, it is unlikely that any of the ensemble members will equal the observation. The lack of correspondence can be accounted for by assigning an error distribution to each ensemble member. In order to do so, one needs to know the appropriate degree of uncertainty. Roulston and Smith's proposal associates uncertainty with the ensemble's best member, where the best member is defined as the one nearest to the observed data in the weather system state space. One application to European ECMWF EPS is shown regarding 4 stations: Tromso, London Heathrow, Frankfurt and Warsaw. Statistical ensembles showed a significant edge in performance over deterministic ensembles. However, substantial residual variability remains, and this cannot be explained by EPS.

The present study starts from the premise that observed data and EPS outputs may be perceived as the components of a statistical model in which observed meteorological values are explained by EPS outputs considered as exogenous variables, as in Raftery et al. 2005. Our model completes the idea of the BMA in two ways. Firstly, it is conceived as a genuinely Bayesian model, and secondly, it deals with non-identifiable EPS outputs such as those present in the European ECMWF EPS, by modelling the second level of the hierarchy in an appropriate manner. In meteorology, observed values do not constitute current inputs for meteorological models and the comparison between observed and predicted values is not as crucial as it can be in other contexts (Gneiting et al., 2007). After integrating out all parameters, our model can be used to obtain the distribution of an as yet unobserved value, conditional on previous observations and forecasts. Our model is not a tool for improving forecasts, but associates probability distributions to EPS outputs, and can explain the residual variability which is not taken into account by an EPS. It obtains truly probabilistic weather predictions which are sharp, calibrated and reliable both for central values and dispersion (see also Hagedorn et al., 2007; Hamill et al., 2007).

The paper is organized as follows. Section 2 illustrates the hierarchical model linking observed data and deterministic forecasts and points the way to obtaining the predictive distribution. Details of the model specification are put off until certain preliminary descriptive analyses of data have been completed at the beginning of Section 3: Section 3.4 completes the model. Section 4 illustrates the results and compares them with the performance of EPS models. Section 5 presents our conclusions.

2 Statistical modelling of EPS: a hierarchical approach

We propose the full Bayesian modelling of EPS output by means of a hierarchical model. For each day t we have (possibly multivari-

ate) observed data y_t, together with K distinct forecast ensemble members X_{tk}. Observed data and EPS outputs can be seen as the components of a statistical model where observed meteorological values are explained by EPS outputs considered as exogenous variables. The principal definitions are summarised in table 1.

We consider forecast scenarios for a particular day as random replications of a single data-generating process. In order to link the ensemble of replicated scenarios to a single observation, we introduce a latent process which randomly selects one scenario from the ensemble. In this way, we model the observed data conditionally on the selection process.

This situation is then translated into a hierarchical model, where the first level concerns observed values as a function of the selected deterministic forecast, while the second level governs the selection of this element.

2.1 Model construction

As we have already said, the EPS output on day t consists of K forecast scenarios:

$$X_t = \{X_{tk}; k = 1, \ldots, K\}$$

These scenarios are unlabelled and for each day there are K new replications, each independent of the previous ones, characterised by a situation of exchangeability.

The first level of the model, i.e. the observation level, where the measurement error occurs, is written as:

$$p(y_t|Z_t, X_t, \theta) = f(X_{t,Z_t}; \theta) \tag{2}$$

where the variable Z_t selects the deterministic forecast for day t. This means that the distribution $f(X_{t,Z_t}; \theta)$ is not a function of all the weather scenarios, but only of the selected ensemble member. In (2) the observed data y_t is a function of just one of the K scenarios on day t.

The scenario selection process is modelled by a latent stochastic process $\{Z_t\}$ which selects, from one day to the next, one scenario from among the K available scenarios, such that the observation on day t is considered as a random fluctuation around the selected weather scenario. As a result of the way in which weather scenarios are generated, there is no correlation in time. We propose to model the latent selection process $\{Z_t\}$ in the second level of the model, i.e. the process level, as i.i.d. with discrete uniform distribution:

$$p(Z_t = k) = \frac{1}{K} \qquad k = 1, 2, \ldots, K \tag{3}$$

such that, for any N-dimensional sequence of times $t_1, t_2, \ldots, t_N$, the following relationship holds:

$$p(Z_{t_1} = k_1, Z_{t_2} = k_2, \ldots, Z_{t_N} = k_N) = \prod_{j=1}^{N} p(Z_{t_j} = k_j)$$

The selected weather scenario can be compactly indicated as a function of Z_t and X_t:

$$x_t = h(Z_t, X_t) = \sum_{k=1}^{K} 1_{\{Z_t = k\}} X_{tk} \tag{4}$$

Conditional on the selection of the weather scenario x_t, data (y_t, X_t) is reduced to the (multivariate) pairs (y_t, x_t). Equation (2) can thus be rewritten with the help of (4) as:

$$p(y_t | Z_t, X_t, \theta) = f(h(Z_t, X_t), \theta) \quad = f(x_t; \theta) \tag{5}$$

In practice, the form of model (5) will depend on the available data set. The details of model specification are therefore deferred to the following Section. A standard choice would be Gaussian regression, with a data-driven specification of the regression function and of variance. Furthermore, if one wants to model a one-dimensional variable in multiple locations, y_t and x_t shall be $S-$variate vectors, and θ shall contain, among other things, variance/covariance parameters

of the distribution of the $S-$variate vector of residuals. Statistical modelling according to (5) has calibration as a natural by-product, a required element which is absent from deterministic models.

We propose a Bayesian solution for model (5)-(3) whereby, after specifying a prior for θ, one finds the posterior distribution $p(\theta|y^{\text{obs}})$, where y^{obs} are the observed data from the training set.

In the following, conditioning on regressor X_t will not be explicitly specified.

2.2 Predictive probability distribution

In the problem in question, we are particularly interested in the posterior predictive probability distribution of the target variable y at a future point in time $T+1$, after observing $y^{\text{obs}} = \{y_t; t = 1, \ldots, T\}$.

By standard Bayesian arguments, we can write:

$$p(y_{T+1}|y^{\text{obs}}) = \int_{\Theta} p(y_{T+1}|\theta)p(\theta|y^{\text{obs}})\mathrm{d}\theta \tag{6}$$

A suitable expression for $p(y_{T+1}|\theta)$ can be obtained from (6) by conditioning w.r.t. Z_{T+1} and applying the total probability law:

$$\begin{aligned} p(y_{T+1}|\theta) &= \sum_{k=1}^{K} p(y_{T+1}|Z_{T+1} = k; \theta) \cdot p(Z_{T+1} = k) \\ &= \frac{1}{K} \sum_{k=1}^{K} p(y_{T+1}|Z_{T+1} = k; \theta) \end{aligned} \tag{7}$$

where the role of the first level of the model written according to (5) is explicitly expressed.

By substituting (7) in (6) we finally get:

$$p(y_{T+1}|y^{\text{obs}}) = \frac{1}{K} \sum_{k=1}^{K} \int_{\Theta} p(y_{T+1}|Z_{T+1} = k; \theta)p(\theta|y^{\text{obs}})\mathrm{d}\theta \tag{8}$$

where the resulting distribution is a mixture of K components with equal weights $1/K$. The integration in (8) may be hard to solve in analytical terms, but can be easily computed using numerical stochastic

Table 1: Notation

variable	description
y_t	observed data on day t
X_{tk}	k-th weather scenario on day t, $k = 1, \ldots, K$
x_t	selected weather scenario on day t

methods. Note that in (8) the parameters vector θ is integrated out from the measurement error model (5), unlike in the BMA solution (1) which replaces θ with a point estimate $\hat{\theta}$ according to an empirical Bayesian approach.

3 Data-driven model completion

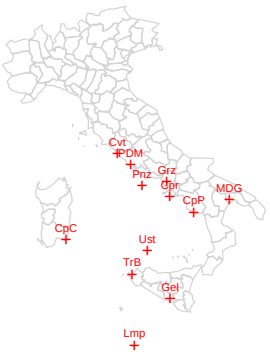
3.1 Forecast data: the COSMO-LEPS system

On the global scale, EPS is a technique commonly used to deal with forecast uncertainty, whereas small-scale use of EPS is still rare. Local meteorological centres generally compute point forecasts exclusively through high resolution Limited Area Models (LAM) applied to global forecasts given by international centres or use multimodel EPS for mimicking global EPS outputs (Krishnamurti et al., 1999; Kharin and Zwiers, 2002).

In 2001, the Emilia Romagna ARPA-SMR (Azienda Regionale Prevenzione e Ambiente - Servizio Meteorologico Regionale) developed LEPS (Local Scale Ensemble Prediction System) which, after an initial experimentation phase, lead to the implementation of COSMO-LEPS (Montani et al., 2001, 2003). The LEPS method employs global-scale ECMWF ensemble forecasts to obtain local-scale, high-resolution ensemble forecasts with relatively low computational costs (Marsigli et al., 2005).

LEPS uses ECMWF ensemble forecasts to obtain a reduced set of local forecasts. In other words, 10 representative members (RM) are selected from the set of 51 ECMWF forecasts by means of an unsupervised cluster analysis. Each RM gives starting and bound

Figure 1: Locations of meteorological stations



conditions for the integration of a high-resolution LAM. So, a corresponding local ensemble member is obtained from each RM.

We consider temperature data at an altitude of *2m* at 12:00, based on deterministic COSMO-LEPS forecasts with a forecast horizon of 24h.

3.2 Observed data

Observed data covers the period from 1/06/2005 to 30/11/2005, and comes from 12 meteorological stations belonging to the Italian Synoptic Network.

The positioning of the selected stations is reported in Fig. 1, while in table 2 geographic details are summarised. In table 3 some descriptive statistics of the observed data are synthesised. From geographic data as well as descriptive statistics, it can be seen how the selected stations are homogeneous w.r.t. orography, distance from sea and overall mean temperature. There are however some differences in the levels of variability, as indicated by the computed standard deviation (sd) and inter-quartile range (IQR) in Gela and Lampedusa, where they are noticeably smaller than in the other stations.

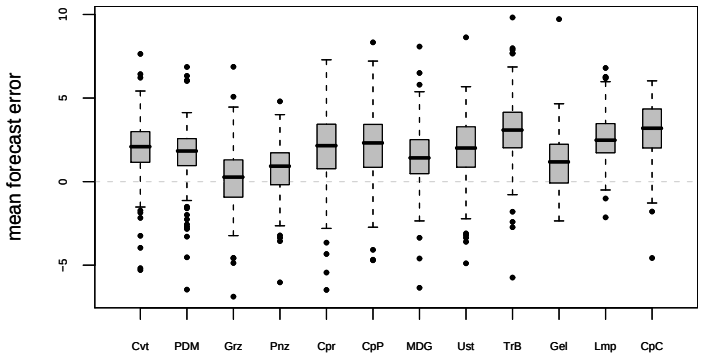
Table 2: Details of the meteorological stations

label	name	lon.	lat.	alt. (m)	missings
Cvt	Civitavecchia	11.83	42.03	3	7
PDM	Pratica Di Mare	12.43	41.65	6	6
Grz	Grazzanise	14.07	41.06	9	6
Pnz	Ponza	12.95	40.92	184	7
Cpr	Capri	14.20	40.55	160	6
CpP	Capo Palinuro	15.28	40.03	184	5
MDG	Marina Di Ginosa	16.88	40.44	2	6
Ust	Ustica	13.18	38.71	250	6
TrB	Trapani Birgi	12.50	37.92	7	5
Gel	Gela	14.22	37.08	11	6
Lmp	Lampedusa	12.60	35.50	16	11
CpC	Capo Carbonara	9.517	39.10	116	11

Table 3: Summary of observed temperature data

station	mean	median	sd	IQR
Cvt	23.12	24.00	4.662	6.00
PDM	23.18	24.60	5.139	6.40
Grz	24.44	26.00	5.913	7.60
Pnz	22.12	24.00	4.791	6.60
Cpr	23.77	25.00	5.655	8.20
CpP	24.27	26.00	5.608	7.65
MDG	24.69	26.00	5.649	7.80
Ust	24.37	24.60	4.71	6.20
TrB	25.74	26.10	4.677	5.60
Gel	24.65	25.00	3.62	3.90
Lmp	25.80	26.20	2.738	2.60
CpC	25.18	26.40	4.618	5.60

Figure 2: Boxplots of mean deterministic forecast errors, by station. The central line is the median.



3.3 Explorative data analysis

Explorative evaluations are useful for suggesting the final details of model building. The differences between observed data and mean ensemble forecasts:

$$\bar{e}_{ts} = y_{ts} - \frac{1}{10}\sum_{k=1}^{10} X_{tsk}$$

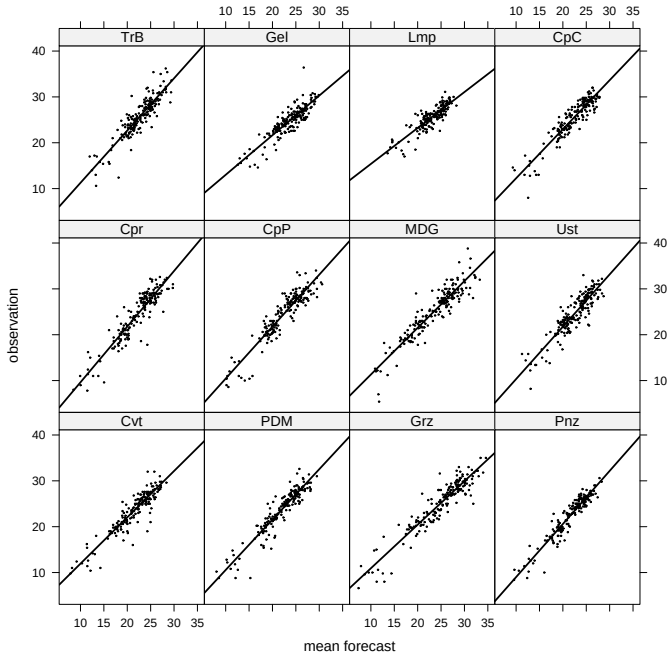
have been considered and reported in Fig. 2 for each meteorological station. Such distributions are roughly symmetric, with comparable ranges. However, there is a positive bias of varying amplitude for all stations, except perhaps for Grazzanise. Further parameters will be introduced during model building in order to account for these differences between stations.

A more detailed view of the relationship between observed data and mean deterministic forecasts is given in Fig. 3. This plot would seem to indicate a linear relationship. On the basis of figure 2, a station-specific intercept should be added, while the slope can be considered approximately constant among stations.

To sum up, our explorative plots suggest:

- the need for a correction for a station-specific positive bias

Figure 3: Mean deterministic forecasts vs observed data, by station



(Fig. 2)

- a linear relationship, with nearly constant slope, between deterministic forecasts and observed data (Fig. 2 and 3)
- approximately constant variance (box-plot width in Fig. 2)

3.4 Statistical model completion

Now we are able to specify model (2) by assuming that x_t and y_t are S-variate vectors:

$$x_t = (x_{t1}, x_{t2}, \ldots, x_{ts}, \ldots, x_{tS})$$
$$y_t = (y_{t1}, y_{t2}, \ldots, y_{ts}, \ldots, y_{tS})$$

and complete the specification of the regression function guided by the exploratory analyses in Section 3.3.

We assume for y_{ts}, conditional on x_{ts} and θ, a Gaussian distribution with expected value constituted by a linear function of the deterministic forecast x_{ts}:

$$y_{ts}|x_{ts}, \theta \sim \mathrm{N}(\alpha_s + \beta x_{ts}, \sigma_y^2) \tag{9}$$

where β is a common slope, α_s are station-specific intercepts such that $\alpha = (\alpha_1, \ldots, \alpha_S)$, and σ_y^2 is a common measurement error variance. Here $\theta_1 = (\alpha, \beta, \sigma_y^2)$, while the y_{ts} scalars are modelled as independent, both in time and space, conditionally on model parameters. In order to complete the model we give a hierarchical structure to the station-specific intercepts:

$$\alpha_s|\alpha_0, \sigma_\alpha^2 \sim \mathrm{N}(\alpha_0, \sigma_\alpha^2) \tag{10}$$

and assign vague priors to β, α_0, σ_y^2 and σ_α^2.

The model parameters consist of the vector $\theta = (\theta_1, \theta_2)$, composed of 12 intercepts, 1 common slope, 1 common variance of the measurement error and 2 hyperparameters $\theta_2 = (\alpha_0, \sigma_\alpha^2)$ for the distribution of intercepts.

4 Results

In order to enable a comparison with COSMO-LEPS ensembles, models have been fitted on a moving window of 30 days, and out-of-sample forecasts simulated for 1 day ahead. This procedure estimates 153 models, corresponding to about 5 months.

For each day, model estimation gives:

- the full joint posterior probability distribution of the parameter vector $p(\theta|y^{\text{obs}})$, from which we can obtain all lower-order marginals
- the predictive distribution function $p(y_{T+1}|y^{\text{obs}})$ computed in (6)

Since we estimate the model for 153 different training sets, the previous estimates are replicated for each target day. This procedure is in keeping with the current work of meteorologists, who look for daily evaluations of forecasts. Unless otherwise stated, for exploratory purposes we have focused on inferences made for the 1^{st} July 2005, which are thus conditional on the $1 - 30$ June 2005.

4.1 The posterior distribution of parameters

One of the results of model estimation is the multivariate posterior probability distribution $p(\theta|y^{\text{obs}})$. Among parameters, 12 are the station-specific intercepts α_i, and we do not discuss them since they are less interesting for the problem under study. In Fig. 4 the marginal and bivariate densities (as isodensity contours in the out-of-diagonal elements) for the remaining 4 parameters are reported.

Bivariate distributions show no pathological posterior dependency between parameters. In fact, pairwise independence holds for almost all parameters, according to the mild shape of the contours, the only exception being the $(\alpha_0, \sigma_\alpha)$ distribution, which shows a form of scale-location dependency.

Parameters β, σ_y and α_0 display (see the diagonal) a symmetric distribution: in the case of β and α_0 this was to be expected, since

Figure 4: Univariate and bivariate parameter densities. Univariate density on the diagonal, bivariate contours out of diagonal.

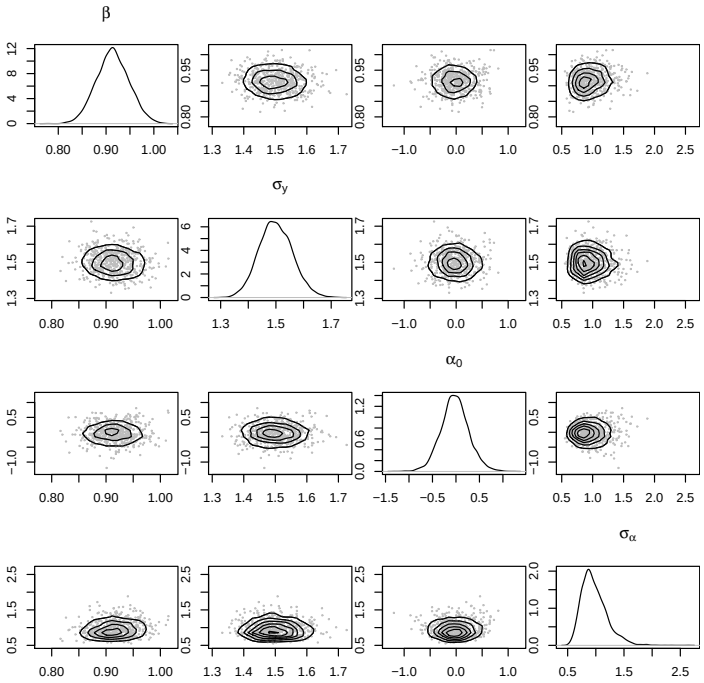


Table 4: Summaries of estimated marginal posterior distributions

parameter	mean	sd	IQR
β	0.916	0.033	0.048
σ_y	1.504	0.063	0.087
α_0	-0.032	0.291	0.348
σ_α	0.971	0.222	0.301

they are location parameters; in the case of σ_y such a shape is likely due to the large amount of independent data that contributed to the estimation (12×30 observed-deterministically forecasted data pairs). The distribution of σ_α is skewed, which is characteristic of dispersion parameters. As regards posterior precision, both β and σ_y display a very sharp marginal distribution, with β roughly ranging from 0.8 to 1.0 and σ_y from 1.3 to 1.7. Note that the distribution of slope β is positioned well below 1. In a certain sense, it is significantly smaller than 1, thus confirming the need for the contraction of the deterministic forecasts in order to obtain improved calibration. The marginal distributions of α_0 and σ_α are much more dispersed than the others, but this is to be expected given that they are second-level parameters.

Table 4 shows some summaries of the marginal posterior distributions. Here the above-mentioned remarks about Fig. 4 are confirmed by the estimated standard deviations (sd) and inter-quartile ranges (IQR).

4.2 Predictive distributions

Fig. 5 plots the predictive probability densities (PD) for the monitoring stations on the 1$^{\text{st}}$ July 2005. The observed value (solid vertical line) and the range of the deterministic ensemble forecast (dotted vertical lines) are shown in each plot. The PD is estimated on data up until the 30$^{\text{th}}$ June. In this way, a fair comparison can be made with the COSMO-LEPS ranges, since we are comparing the statistical forecast with data which did not contribute to model estimation.

Table 5: Observed and mean predicted values for the 1st July 2005

station	observed	EPS mean	PDF mean
Cvt	27.6	26.5	28.5
PDM	28.2	26.3	28.3
Grz	28.8	28.6	28.9
Pnz	25.8	24.6	25.9
Cpr	32.0	25.3	28.3
CpP	29.0	26.6	28.5
MDG	32.0	26.2	28.2
Ust	28.4	25.0	27.7
TrB	29.0	24.8	27.3
Gel	27.6	27.2	28.2
Lmp	28.9	24.9	26.7
CpC	29.4	25.3	28.3

Table 5 compares observed data and mean predicted values.

The observed values at the various stations may or may not fall within the COSMO-LEPS range, hence the significant contribution made by the statistical model to deal with the weather forecasting problem. In fact, when the observed data falls within the EPS range (`Grz` and `Gel`), the PD mean also lies within that range. When the observed data falls outside the said range, the PD tends to indicate those values that fall between the deterministic COSMO-LEPS range and the actually observed data as the most probable (`Cpr`, `CpP`, `MDG`, `Ust`, `TrB`, `Lmp` and `CpC`). In some cases (`Cvt`, `PDM` and `Pnz`), the PD emphasises the differences noticed between the COSMO-LEPS range and the observed value, but with more weight given to values closer to the observed value. In all cases, the overall tendency sees the PD assume the highest values that are closest to the observed value (but that have not been used for model estimation).

Note that the forecast calibration is not an explicit target of model estimation, but is the most important by-product of the fitting process, and one that does not require any form of human tuning.

In Fig. 6 we report the set of daily Ustica PD computed for July-August. The darker area indicates the central 50% Highest Density

Figure 5: Observed data, the COSMO-LEPS range and model forecast for the 1st July 2005

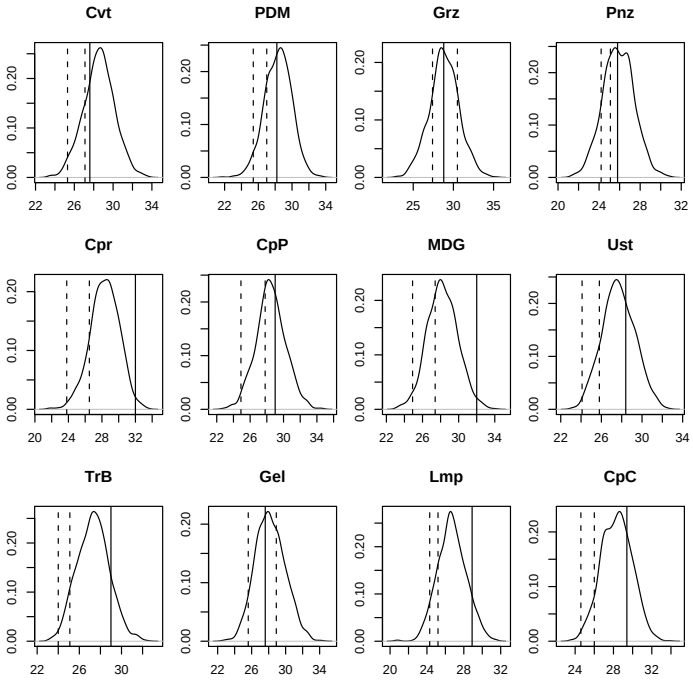
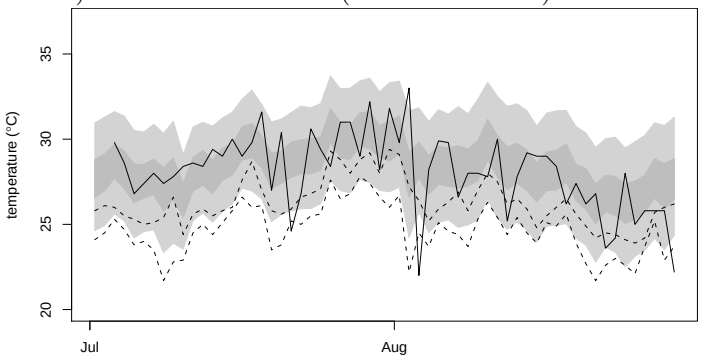


Figure 6: July-August HDI (coloured area), COSMO-LEPS range (dotted lines) and observed data (continuous line) for Ustica



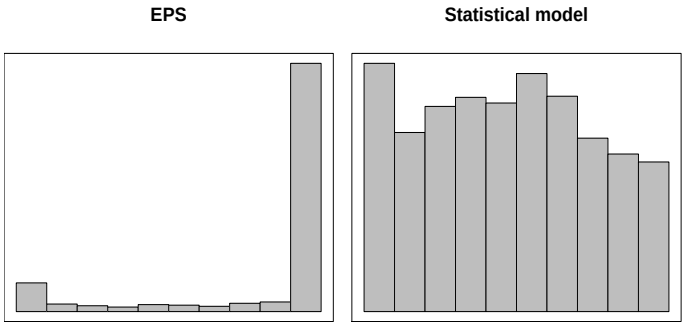
Interval (HDI), while the lighter area represents the 90% HDI. The dotted lines indicate the EPS range, while the continuous line represents the observed data. This plot once again reveals the shift of the model-based PD towards the observed data for the 2 months in question, as well as the good performance of the PD in relation to both the central value and dispersion.

4.2.1 Verification Rank Histograms

In ensemble forecasting practice, certain standard diagnostic measures are employed to evaluate the performances of EPS. These measures can be applied to the output of our statistical model, thus enabling a direct comparison to be made with the COSMO-LEPS result.

A common diagnostic device is the Verification Rank Histogram (VRH; see Talagrand et al., 1997). The VRH is the histogram of frequencies of the rank of the observed data within the forecast ensemble. A good EPS should have a uniform distribution, meaning exchangeability between deterministic predictions and observed data (Buizza, 1997; Buizza et al., 2005; Gneiting et al., 2007). Other shapes may indicate under/overdispersion and bias. In particular, a marked U-shaped distribution means that observed values tend to

Figure 7: Overall VRH



be systematically outside (above or below) the extremes of the EPS range. A concave shape indicates that the forecasting range is larger than what is actually needed.

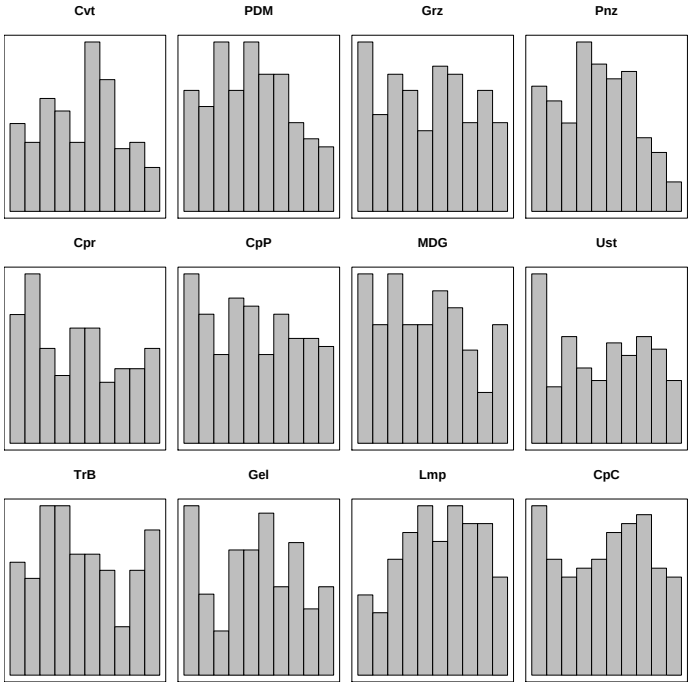
Fig. 7 shows the VRH for the full dataset, both for the statistical model and for the deterministic EPS. It shows a clear lack of calibration in the case of deterministic EPS, while the results coming from the statistical models present a rather uniform distribution, indicating a good mix of model-based out-of-sample forecasts and actually observed data.

If we compute station-specific VRHs, we do not obtain as good results as for the overall histogram. Less data is available for each VRH, and the overall histogram is an average of the local evaluations. In each station-specific VRH, the shape is however far from the U-shape displayed by the global COSMO-LEPS VRH (see Fig. 8).

4.3 MCMC diagnostics

The model built in Sections 2 and 3 cannot be estimated analytically, but requires the use of numerical methods. MonteCarlo Markov Chains (MCMC) are a natural choice for the estimation of hierarchical models. MCMC can be problematic with respect to chain convergence and mixing (Robert and Casella, 1999). However, with regard to the problem in question, thanks to the adoption of conjugate priors

Figure 8: Station-specific statistical model VRH



for all parameters, which give standard full conditional distributions, as in Gelman 2004, genuine Gibbs sampling can be used. The Gibbs sampler performed well in this particular case.

So, for each daily model estimation, 5000 MCMC iterations were kept after discarding the first 1000. From each estimated model, 1000 scenarios were simulated from the predictive distribution. Chains revealed fast convergence and good mixing properties, with no need for ad-hoc fine tuning.

For the sake of example, we show some MCMC diagnostics for the model fitted in the period 1-30 June 2005 and used on the 1st July 2005. Fig. 9 shows MCMC traces, plotting sampled values vs. iteration number. These plots show a rapid convergence towards the target stationary distribution, with no real need to cut out any starting transient. Chains mixing is assessed here by means of a linear autocorrelation index. Fig. 10 plots sample autocorrelations at 36 different lags for the 4 main model parameters. Autocorrelations are almost null at all lags for α_0, σ_y and σ_α, and negligible at lags > 3 for β. This guarantees that the sampling process is highly efficient, so that there is no need for many replicates in order to reliably estimate parameters distribution.

5 Conclusions

EPS try to account for uncertainty due to partial knowledge of starting conditions, but without explaining residual variability. This study presents a genuinely Bayesian framework designed to deal with EPS constituted by unidentifiable members. An explicit link between EPS output and observed data is posited by introducing a latent selection process. We illustrate one application to small-scale temperature data for 12 meteorological stations. The model's output is the full multivariate posterior probability distribution of the set of parameters characterizing the model, from which the following marginal, as well as bivariate, syntheses can be computed: central values, standard deviations, cross-correlations, credibility intervals, probability of falling in fixed ranges, etc. Our probabilistic model also allows us

Figure 9: MCMC traces

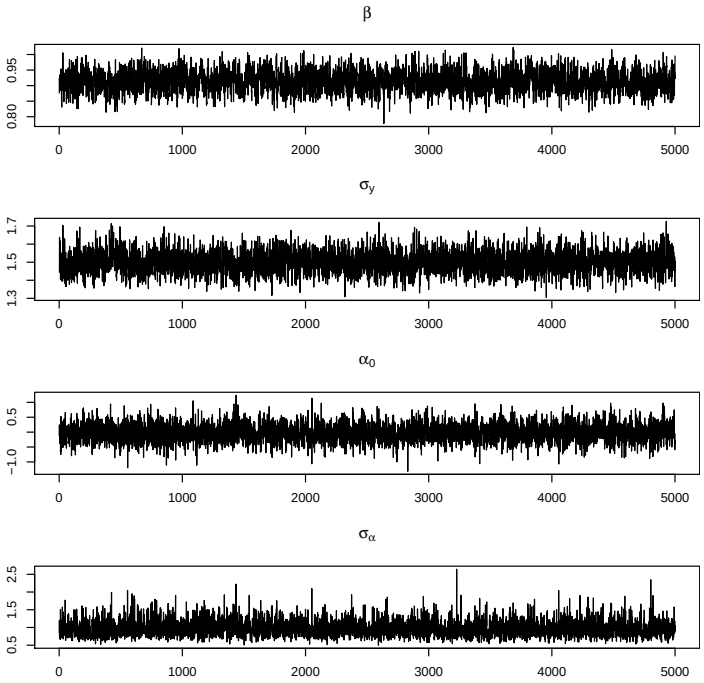
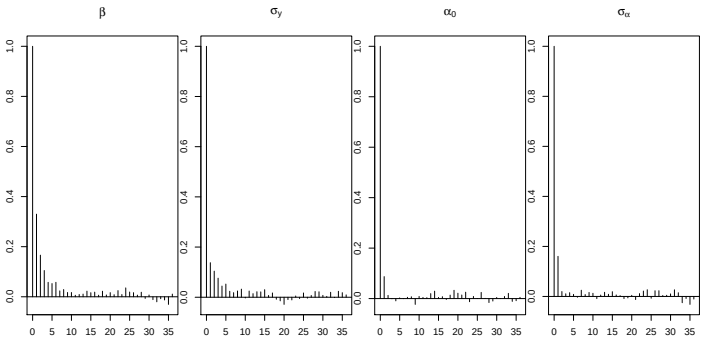


Figure 10: MCMC autocorrelations



to compute reliable, informative predictive distributions.

Since a Bayesian model produces a wealth of information, its results can be interpreted in a variety of different ways. One way of decoding such results is the combined use of synthetic values of the posterior with the predictive distributions. The description of our model's results in fact makes use of several such syntheses, combining them in a flexible way, thus showing how a statistical model can help understand and improve deterministic models. In more specific terms, the predictive distribution obtained by the statistical model tends to be nearer to the observed values than does the COSMO-LEPS output.

Unfortunately, statistical modelling does not reduce the variability of results. Perhaps a more sophisticated form of modelling could represent a step in this direction. Indeed, statistical models tend to be rather simplistic, and their added value is constituted by their ability to bring together the various different components of the problem in hand in a suitable, consistent manner. Among such components, account is taken of the sources of variability in the right place. Finally, one natural by-product of statistical models is calibration, which is often required.

References

Buizza, R. (1997). Potential forecast skill of ensemble prediction and spread and skill distributions of the ecmwf ensemble prediction system. *Mon. Wea. Rev. 125*, 99–119.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu (2005). A comparison of the ecmwf, msc and ncep global ensemble prediction systems. *Mon. Wea. Rev. 133*, 1076–1097.

Gelman, A. (2004). *Bayesian Data Analysis*. CRC Press.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc. B 69*, 243–268.

Hagedorn, R., T. Hamill, and J. Whitaker (2007). Probabilistic forecast calibration using ecmwf and gfs ensemble reforecasts. part i: 2-meter temperatures. *Monthly Weather Review*. Accepted for pubblication.

Hamill, T., R. Hagedorn, and J. Whitaker (2007). Probabilistic forecast calibration using ecmwf and gfs ensemble reforecasts. part ii: Precipitation. *Monthly Weather Review*. Accepted for pubblication.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Kharin, V. V. and F. W. Zwiers (2002). Climate predictions with multimodel ensembles. *J. Climate 15*, 793–799.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendan (1999). Improved weather and seasonal climate forecasts from multimodel superensembles. *Science 258*, 1548–1550.

Leith, C. E. (1974). Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev. 102*, 409–418.

Marsigli, C., F. Boccanera, A. Montani, and T. Paccagnella (2005). The cosmo-leps memoscale ensemble system: validation of the methodology and verification. *Nonlinear Processes in Geophysics* (12), 527–536.

Montani, A., M. Capaldo, D. Cesari, C. Marsigli, U. Modigliani, F. Nerozzi, T. Paccagnella, and S. Tibaldi (1996). The ecmwf ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.* (122), 73–119.

Montani, A., M. Capaldo, D. Cesari, C. Marsigli, U. Modigliani, F. Nerozzi, T. Paccagnella, and S. Tibaldi (2001). A strategy for

high-resolution ensemble prediction, part i: Definition of representative members and global model experiments. *Quart. J. Roy. Meteor. Soc.* (127), 2069–2094.

Montani, A., M. Capaldo, D. Cesari, C. Marsigli, U. Modigliani, F. Nerozzi, T. Paccagnella, and S. Tibaldi (2003). Operational limited-area ensemble forecasts based on the lokal modell. *ECMWF Newsletter* (98), 2–7.

Raftery, A. E., T. Gneiting, F. Balabdaqui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *American Meteorological Society 133*, 1155–1173.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York: Springer.

Roulston, M. S. and L. A. Smith (2003). Combining dynamical and statistical ensembles. *Tellus 55A*, 16–30.

Talagrand, O., R. Vautard, and B. Strauss (1997). Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability*, 1–25.

Toth, Z., Y. Zhu, and T. Marchock (2001). The use of ensembles to identify forecasts with small and large uncertanity. *Wea. Forecasting 16*, 463–477.